AFRL-RX-WP-JA-2016-0304

BEST PRACTICES FOR EVALUATING THE CAPABILITY OF NONDESTRUCTIVE EVALUATION (NDE) AND STRUCTURAL HEALTH MONITORING (SHM) TECHNIQUES FOR DAMAGE CHARACTERIZATION (POSTPRINT)

Eric A. Lindgren
AFRL/RX

John C. Aldrin
Computational Tools

Charles Annis
Statistical Engineering

Harold A. Sabbagh
Victor Technologies

6 October 2015
Interim Report

(STINFO COPY)
AIR FORCE RESEARCH LABORATORY
MATERIALS AND MANUFACTURING DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7750
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
6 October 2015	Interim	6 May 2010 – 6 September 2015

4. TITLE AND SUBTITLE
BEST PRACTICES FOR EVALUATING THE CAPABILITY OF NONDESTRUCTIVE EVALUATION (NDE) AND STRUCTURAL HEALTH MONITORING (SHM) TECHNIQUES FOR DAMAGE CHARACTERIZATION (POSTPRINT)

5a. CONTRACT NUMBER
FA8650-10-D-5210

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
1) Eric A. Lindgren – AFRL/RX
2) John C. Aldrin – Computational Tools
(continued on page 2)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
X0N6

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
1) AFRL/RX
Wright-Patterson AFB, OH 45433
2) Computational Tools
4275 Chatham Avenue
Gurnee, IL 60031
(continued on page 2)

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
Materials and Manufacturing Directorate
Wright-Patterson Air Force Base, OH 45433-7750
Air Force Materiel Command
United States Air Force

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
AFRL/RXCA

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RX-WP-JA-2016-0304

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
PA Case Number: 88ABW-2015-4793; Clearance Date: 6 Oct 2015. This document contains color. Journal article published in AIP Conference Proceedings, Vol. 1706, No. 1, 10 Feb 2016.
The final publication is available at http://dx.doi.org/10.1063/1.4940646

14. ABSTRACT ***(Maximum 200 words)***
A comprehensive approach to NDE and SHM characterization error (CE) evaluation is presented that follows the framework of the 'ahat-versus-a' regression analysis for POD assessment. Characterization capability evaluation is typically more complex with respect to current POD evaluations and thus requires engineering and statistical expertise in the model-building process to ensure all key effects and interactions are addressed. Justifying the statistical model choice with underlying assumptions is key. Several sizing case studies are presented with detailed evaluations of the most appropriate statistical model for each data set. The use of a model-assisted approach is introduced to help assess the reliability of NDE and SHM characterization capability under a wide range of part, environmental and damage conditions. Best practices of using models are presented for both an eddy current NDE sizing and vibration-based SHM case studies. The results of these studies highlight the general protocol feasibility, emphasize the importance of evaluating key application characteristics prior to the study, and demonstrate an approach to quantify the role of varying SHM sensor durability and environmental conditions on characterization performance.

15. SUBJECT TERMS
Nondestructive evaluation (NDE); Structural Health Monitoring (SHM); POD

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Charles Buynak
Unclassified	Unclassified	Unclassified	SAR	13	19b. TELEPHONE NUMBER *(Include Area Code)* (937) 255-9807

REPORT DOCUMENTATION PAGE Cont'd

6. AUTHOR(S)

3) Charles Annis - Statistical Engineering

4) Harold A. Sabbagh - Victor Technologies

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

3) Statistical Engineering, 7243 Oxford Court,
Palm Beach Garden, FL 33418

4) Victor Technologies LLC, 2609 S Spicewood Ln,
Bloomington, IN 47401

Best Practices for Evaluating the Capability of Nondestructive Evaluation (NDE) and Structural Health Monitoring (SHM) Techniques for Damage Characterization

John C. Aldrin[2, a)], Charles Annis[3, b)], Harold A. Sabbagh[4, c)], and Eric A. Lindgren[1, d)]

[1]*Air Force Research Laboratory (AFRL/RXCA), Wright-Patterson AFB, OH 45433*
[2]*Computational Tools, 4275 Chatham Avenue, Gurnee, IL 60031*
[3]*Statistical Engineering, Palm Beach Garden, FL 33418*
[4]*Victor Technologies LLC, Bloomington, IN 47401*

[a)] aldrin@computationaltools.com
[b)] has@sabbagh.com
[c)] charles.annis@statisticalengineering.com
[d)] eric.lindgren@us.af.mil

Abstract. A comprehensive approach to NDE and SHM characterization error (CE) evaluation is presented that follows the framework of the 'ahat-versus-a' regression analysis for POD assessment. Characterization capability evaluation is typically more complex with respect to current POD evaluations and thus requires engineering and statistical expertise in the model-building process to ensure all key effects and interactions are addressed. Justifying the statistical model choice with underlying assumptions is key. Several sizing case studies are presented with detailed evaluations of the most appropriate statistical model for each data set. The use of a model-assisted approach is introduced to help assess the reliability of NDE and SHM characterization capability under a wide range of part, environmental and damage conditions. Best practices of using models are presented for both an eddy current NDE sizing and vibration-based SHM case studies. The results of these studies highlight the general protocol feasibility, emphasize the importance of evaluating key application characteristics prior to the study, and demonstrate an approach to quantify the role of varying SHM sensor durability and environmental conditions on characterization performance.

INTRODUCTION

The current U.S. Air Force practice for maintaining aircraft structures follows the Aircraft Structural Integrity Program (ASIP) methods, as documented in MIL-STD-1530C [1]. Following this damage tolerance approach, the periodic inspection of structures is performed using validated nondestructive evaluation (NDE) techniques. In addition, there is a significant interest to certify the capability of nondestructive evaluation techniques to perform damage characterization. As the maintenance of the structural components of aircraft moves from time-based maintenance to condition-based maintenance, there is a need for innovative methods to not simply detect damage, but to completely characterize it in structural components [2]. For example, accurate knowledge of crack location and size would improve decision-making concerning maintenance actions, reduce unnecessary teardowns, minimize maintenance induced damage, and provide key information for prognostics programs. Concerning emerging structural health monitoring (SHM) techniques, the necessary component of any reliability demonstration is ensuring that damage characterization errors are well understood and within acceptable ranges. This information is critical in order to determine the real benefit of an SHM technique on the economic service life and risk for an aircraft [3].

42nd Annual Review of Progress in Quantitative Nondestructive Evaluation
AIP Conf. Proc. 1706, 200002-1–200002-10; doi: 10.1063/1.4940646

Probability of detection (POD) evaluation procedures have been developed to validate the reliability of NDI techniques and used by the USAF in support of ASIP [4]. Building on prior work, the goal of this effort is to develop a procedure with statistical tools to properly evaluate NDE characterization techniques for accuracy in sizing and/or locating damage. A comprehensive approach to NDE and SHM characterization error (CE) evaluation is presented that follows the framework of the 'ahat-versus-a' regression analysis for POD assessment. A key point is that reliability evaluation is likely more complex with respect to current POD evaluations and indicates the importance of engineering and statistical expertise in the model-building process to ensure all key effects and interactions are addressed. Several sizing case studies are presented with detailed evaluations of the most appropriate statistical model for each set. Lastly, a discussion is presented on the use of a model-assisted approach to assess the reliability of NDE and SHM characterization capability. Best practices of using models are presented for both an eddy current NDE sizing and vibration-based SHM case studies. The results of these studies highlight the general protocol feasibility, emphasize the importance of evaluating key application characteristics prior to the study, and demonstrate an approach to quantify the role of varying SHM sensor durability and environmental conditions on characterization performance.

EVALUATION OF NDE CHARACTERIZATION CAPABILITY

MIL-HDBK-1823A provides procedures and guidance on statistical analysis for performing a probability of detection (POD) evaluation to validate the reliability of NDI techniques [4]. Building on this prior work, the goal of recent work has been to develop a procedure with statistical tools to properly evaluate NDE characterization techniques for sizing and/or locating damage. There have been some recent efforts to define and demonstrate a complete process for evaluating sizing capability, specifically addressing discontinuities in welds and corrosion in aircraft structures. (A survey of this prior work is presented in [5].) However, there are some outstanding issues with the current practice for the quantitative evaluation of sizing capability with respect to NDE technique evaluation. One metric frequently cited is the calculation of the 95% *safety limit against undersizing* (LUS) bound for quantifying sizing performance for discontinuities in welds [6]. However, there are some important assumptions, such as linearity in the response and constant variance with changes in flaw size that should be addressed before using this metric. In addition, the simplistic character of the bound from a least squares fit does not ensure it will adequately address the true variation of the bound with the varying distribution of discontinuities and limited sample numbers. A more rigorous process is needed to ensure that the bounds on sizing performance being reported from any study are valid.

From the perspective of quantifying the reliability of NDE and SHM systems, there is a need to evaluate the relationship between the accuracy in estimating the damage or material state estimates ($\hat{a}$) with respect to the actual condition (a). An evaluation of the characterization error (CE), $\hat{e} = \hat{a} - a$, for all critical location and sizing estimates is necessary. Characterization error with prediction bounds is the metric that will be used for condition-based maintenance and prognosis programs to help evaluate remaining life and determine necessary maintenance actions. This problem of evaluating characterization error with prediction bounds as a function of a critical parameter such as flaw size is shown in Fig. 1(a). This evaluation is generally similar to the current procedure found in MIL-HDBK-1823A [4] for the evaluation of the relationship between an NDE measurement ($\hat{a}$) and a critical flaw size (a) as shown in Fig. 1(b). Thus, the proposed foundation for the experimental-based CE procedure including will be MIL-HDBK-1823A [5, 7].

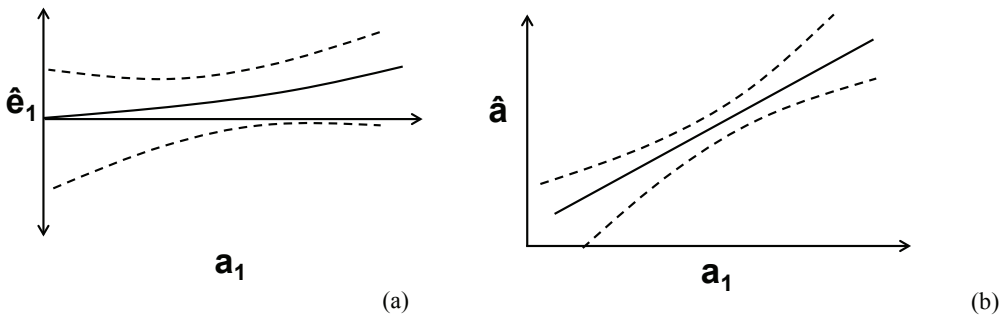

FIGURE 1. Evaluation of (a) characterization error (CE) ($\hat{e}_j$) with respect to damage conditions (a_k) and (b) relationship between discontinuity size (a) and measurement response (â) given by an â-vs-a probability of detection (POD) model.

IMPLICIT STATISTICAL ASSUMPTIONS IN REGRESSION ANALYSIS

Ordinary Least-Squares (OLS) linear regression analysis relies on assumptions concerning the relationship between reality and the process being modeled. Perhaps the most obvious assumption is, "The model must look like the data". While this may be self-evident, checking to see if the assumption holds is less so in practice. There are five other implicit assumptions that must be satisfied for the resulting parameter estimates to be useful:

1. The response must be continuous and observable.
2. The model must be linear in the parameters.
3. The variance must be *homoscedastic* (uniform variance)
4. The observations must be uncorrelated (with respect to time, space, or both).
5. The errors must be Normal.

If any of these assumed conditions are not met, the resulting analysis will be wrong, even though that fact may be far from obvious. The assumptions also hold for the method of Maximum-Likelihood Estimation (MLE), frequently used in POD evaluation of NDE and SHM capability. For more details on the implicit assumption in regression analysis, see ref. [8].

The importance of verifying assumptions of the statistical model is demonstrated in Fig. 2. An example is first presented in Fig. 2(a), where all response points are observed. This is not always the case. For example, sometimes the response is below some noise threshold, or above some saturation value. In that case, it is *censored*. Since these data points are unknown (other than being below some noise or above some saturation level), it is obviously not possible to compute the difference (error) between the observation and the model. Thus, finding a summed squared error is not possible. However, they should not be ignored, which means throwing away useful information. Figure 2(b) illustrates that the OLS parameter estimates based on replacing an observation with its censoring value results in an erroneous, anticonservative, POD vs size model. This is clearly unacceptable. Fundamentally, censored data OLS regression is untenable. However, it is possible to frame the problem in terms of in terms of likelihood estimation. With likelihood, there are a collection of observations and the objective is to evaluate the likely mean and standard deviation of the data. How can censored observations be address? We don't know X, only that it is smaller than, or larger than, some censoring value. We also don't know the ordinate. Since X could be anything in the censored region, one can define the likelihood of a censored observation as *all* of them, *i.e.* the integral of the probability density below, or above, the censoring value. Then, the optimization problem is solved. Rather than minimizing the summed error, the likelihood can be maximized through an optimization scheme. When the data are not censored, maximum likelihood estimators are *exactly* equal to OLS estimators, so we don't need to jettison 200 years of OLS experience to use the MLE criterion. However, OLS is powerless to deal with censoring, but likelihood handles censored data easily. In Fig. 2(c), results are presented using MLE with the regression data in Fig. 2(b), which shows the correct censored regression fit as compared with the OLS fit of all the data in Fig. 2(a). It is not perfect, but it is far superior to using the wrong model. For more details on the theory on MLE, see ref. [8].

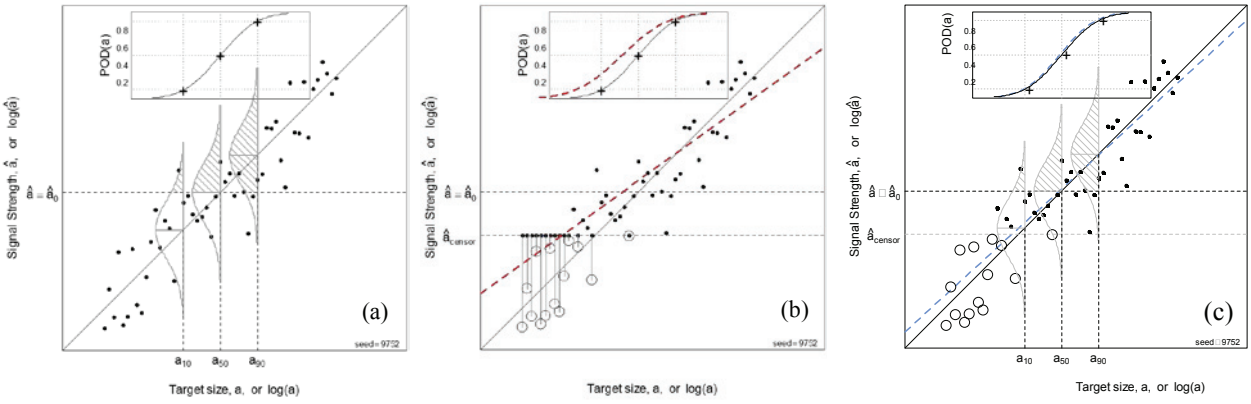

FIGURE 2. (a) OLS requires all responses to be observable. (b) Replacing censored values with the censoring value skews the result anticonservatively. (c) Censored regression using MLE (blue dashed line) correctly accounts for observations with actual responses obscured by background noise and thus censored

It is important to consider that NDE and SHM characterization error evaluation will likely be more complex with respect to current POD evaluations and indicates the importance of engineering and statistical expertise in the model-building process to ensure all key effects and interactions are addressed. As with any POD evaluation study, it is important to plot the data and determine the best statistical model to apply to the evaluation. Before immediately fitting a statistical model to the data, an intermediate step is needed to evaluate the presence and frequency of several possible classes of poor characterization results due to: (1) weak signals from small discontinuities masked by measurement noise, (2) saturated signals or conditions exceeding the inversion algorithm design, (3) ill-posed inversion problems leading to clustering in local minima, (4) random poor characterization performance due to a process failure independent of flaw size. For more information on regression models for evaluating characterization error, see ref. [5, 7-8].

EDDY CURRENT NDE CASE STUDY

An eddy current crack sizing case study is presented to highlight examples of some of these complex characteristics of analyzing sizing error data [5]. First, weak signals from small discontinuities masked by measurement noise will undoubtedly be difficult to size. Such data should be practically removed from the evaluation and clearly reported. For this case study, indications 'A' and 'B' shown in Fig. 3(a) appear to be associated with very small flaw signals. In a POD evaluation, this characteristic in the data is called 'left censoring'. Statistical methods in POD evaluation to address left-censored data can be applied for CE evaluation. Second, strong signals from very large discontinuities can also present several issues for NDE sizing techniques. In practice, it is possible that operators do not require exact sizing when certain flaws become exceptionally large. For example, if a crack above a certain large size is found, the part will simply be replaced. In POD evaluation, this characteristic of saturated data is called 'right censoring'. Statistical methods in POD evaluation to address right-censored data can be directly applied to CE evaluation.

At this stage, the characterization error results are plotted with 'left' and 'right' censoring as shown in Fig. 3(b)-(d). For the BHEC inversion case study problem, estimated crack sizes with lengths below 50 mils and/or crack depths estimated below 18 mils, and crack depths above 100 mils were removed from the data set. Next, it is important to check the data and evaluate where secondary clusters or trends in the data are present. Certain challenging inverse

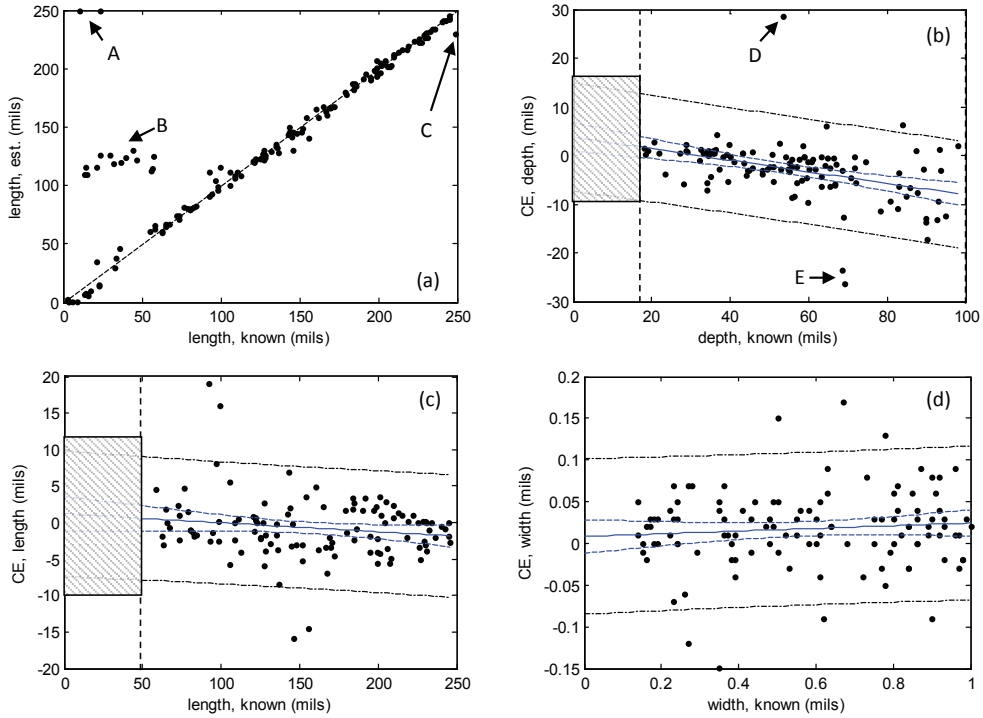

FIGURE 3. Sizing results for BHEC (a) crack length. Characterization error for censored inversion results for (b) crack depth, (c) crack length and (d) crack width. Plots include a linear model fit (solid line) with confidence bounds (dashed line) and corresponding prediction bounds (dash-dot line).

problems are known to have issues with ill-posededness, where the results are prone to clustering at a poor solution due to getting caught in local minima. Case 'B' is an example where signals near the noise threshold are difficult to size and then being mistakenly called in a region that underestimates depth while overestimating length. If such data are not 'left' censored or 'right' censored, then more sophisticated statistical models may be needed to address such results. Lastly, for some NDE inspections that are highly dependent upon human factors, random poor characterization performance can sometimes arise. Such instances are assumed to be due to a process failure that is considered independent of flaw size. This condition is referred to as *random missed call rate* in a POD evaluation and certain statistical methods have been developed to evaluate the rate of random missed calls during a POD study. For the characterization error results for crack depth shown in Fig. 3(b), there are a few indications, 'D' and 'E', that appear to be significantly outside the main scatter of data. Ideally, further analysis of the source for such cases is needed to determine if exceptionally poor data are the source of the 'outlier' or if it is simply due to having limited data samples and should be included in the primary statistical model fit. For this example, indications 'D' and 'E' will be included in the primary statistical model fit. For the characterization error model for the BHEC sizing problem, a linear model with covariance was evaluated for the results in Fig. 3(b)-3(d) based on the following linear model relationship:

$$\mathbf{CE}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_\theta + \boldsymbol{\beta}_1\boldsymbol{\theta}, \mathbf{C}_\theta). \tag{1}$$

The fit was performed using maximum likelihood estimation in R statistical software. The characterization error model plots in Fig. 3 include a linear model fit with 95% confidence bounds (in blue) and 95% prediction bounds (in black dash-dot). In general, the linear model fit appears to be adequate for the censored data set presented here. However, even with this generally well behaved data set, there is a significant change in the lower bound for error in the crack length and depth estimates as a function of changing varying size. This demonstrates the need to take care when attempting to report a single value that defines the entire lower bound for the 'safety limit against undersizing'. Operators should not ever mandate generating such simple metrics when they are often not appropriate for the data.

BEST PRACTICES FOR SHM CAPABILITY EVALUATION

The successful deployment of systems for health monitoring of structures depends on appropriate verification and validation (V&V) of these SHM systems. The V&V method must explicitly evaluate all aspects of the SHM system that can affect its capability to detect, localize, or characterize damage. Moreover, it must evaluate the effects that usage and environmental conditions have on these capabilities over time. As SHM methods depending on permanent, on-board mounted damage sensing systems continue to be proposed and developed for complementing ground-based NDE inspections for aircraft structural integrity purposes, it is necessary that the reliability of these damage sensing systems be assessed with a rigor that is suitable and sufficient for the function that they are expected to perform within the ASIP methodology [13]. For damage detection, this necessarily results in the need for a POD determination. For localization and characterization, characterization error metrics and their evaluation process have been presented and demonstrated [5, 7-8].

Recent work within the structural health monitoring community has strived to define the necessary requirements for SHM certification [9-11] and demonstrate SHM capability through POD evaluation studies [12-25]. In most cases, these POD studies are quite limited in scope, where the POD results only apply to the constraints and assumptions of each study. However, important insight has been identified through these efforts. A recent workshop on SHM Reliability was held in 2015 to discuss progress and address remaining challenges [26]. There does appear to the consensus within the community that if SHM is used to mitigate life-cycle risk of an airframe under the framework of ASIP, a rigorous capability study following the spirit of MIL-HDBK-1823A for POD evaluation is necessary. However, many challenges remain, especially for the certification of global SHM systems. Key issues can be summarized as follows:

- Extreme environmental operation conditions which can impact sensor performance and damage observability,
- Sensor degradation and issues with on-board sensor calibration over time,
- Sensitivity to critical damage while minimizing false calls due to varying non-damage conditions, (For SHM, environmental conditions, material properties, and part geometry will all exhibit some degree of variation for each instance of monitoring. Detecting damage in the presence of non-critical changes to the structure is necessary for SHM systems to provide any value to ASIP.)

- Sensitivity to damage as a function of both damage location and sensor placement. (It has been shown in several local and global SHM reliability studies [15-16, 23] that one cannot typically assume independence for either sensor location or defect location, and generate a single POD curve covering all inspection scenarios. All statistical model assumptions must be properly validated in any POD studies.)

The perspective of the authors of this paper is that the procedures in references [4, 12-17] provide an excellent framework for SHM capability evaluation and address many of the above challenges. A summary of this framework is as follows:

- This protocol includes four critical components: (1) a procedure to identify the critical factors impacting SHM system performance; (2) a multistage or hierarchical approach to SHM system validation; (3) a model-assisted evaluation process to address the wide range of expected damage conditions that cannot be experimentally tested; and (4) POD, probability of false call (POFC) and probability of random missed call (POMC) evaluations with confidence bounds estimation and uncertainty analysis for damage. [14]
- The multistage evaluation approach (2) includes (a) laboratory testing of relevant flaws, (b) laboratory sub-component testing including environmental and loading conditions, (c) a system level life-testing (full-scale fatigue testing if feasible), (d) on-structure demonstration, and (e) final system verification. [14,17]
- The following opportunities in the POD model evaluation process (3 and 4) have the potential to impact sample and testing requirements: (a) careful *model factor selection* addressing system variation, (b) *physics-based model calibration* including uncertainty bounds assessment for the specific inspections of interest, (c) controlled *physics-based model validation* to ensure the model is valid over desired range of application, (d) *evaluation of POD using two-level analysis* to address input *parameter variability with uncertainty bounds*, (e) *integration of experimental data* generated from a *designed experiment* using a *Bayesian framework* to revise the prior distributions of inputs and achieve new posterior distributions [28], and (f) *inverse methods* to ideally address all uncontrolled parameter variations in the measurement. [14-16]

However, the devil is in the details. More rigorous demonstrations are needed for the most-promising SHM technologies to work through these evaluation procedures and address outstanding challenges, especially performing evaluations within a limited budget. One case study for a vibration-based SHM system is given below.

CASE STUDY FOR SHM CAPABILITY EVALUATION

The example used for this initial demonstration of the protocol is a system for detecting the presence of damage using permanently mounted transducers. A test article representing an aircraft structure of medium complexity was designed and built. The test article consists of three plates connected by two lap joints with fasteners. In addition, a fixture was built for supporting the test article. Fatigue crack damage around the fastener holes can be simulated by manually created thin cuts at selected locations. The test fixture design provides the capability to vary critical parameters of the system with a focus on force loading boundary conditions, joint fastener torque conditions, and temperature. The initial demonstration on this test article and fixture uses a vibration based damage detection method. Variations in operational temperature were simulated by testing the system inside a carefully controlled Thermotron SE-1200 environmental chamber. More details on the hardware, DAQ system and damage detection algorithm used in the experiment can be found in [15-16].

Key Factor Evaluation Studies

Following the protocol introduced in reference [14], prior to designing the validation test matrix, the following factors were assessed through controlled studies: (a) mass loading and unloading, (b) fastener torque, (c) boundary condition variation, (d) temperature variation and temperature gradients, (e) sensor bond quality, (f) ambient noise, and (g) flaw growth. More details on the factor evaluation studies are presented in reference [15-16]. Some key results highlighting the need for this task are provided below.

Thermal loading studies were performed by varying the ambient temperature from -20°F to 150°F. During this study, the thermal capacity of the panel end conditions fixtures was found to produce significant thermal gradients across the test article. During heating and cooling periods, temperature gradients as high as 45°F across the test specimen were observed. For validation studies, an estimate of expected gradients ‘in the field’ is needed. An

assumption was made for the validation study that temperature gradients in the region of interest will be primarily limited to +/- 10°F.

Failure of accelerometer bonding was observed several times during thermal testing. These failures occurred during the prolonged high temperature runs at 150°F. While vibration-based damage detection systems are proposed as global methods with some sensor redundancy, relying on a single reference sensor will result, upon bond degradation, in either highly degraded performance and/or complete failure of the damage detection system. Sensor and sensor bonding reliability must be accurately assessed as part of a validation study.

An initial study on the effect of damage growth was performed to ensure adequate sensitivity during the final validation study. Customized XActo™ blades were used to make cuts in the aluminum plates. Cuts were initially made at 0.063" (1/16") increments up to 0.63". For the first series of cuts up to 0.25", sensitivity to notch length increases was observed, but the trend was small relative to noise, and not quite linear. Greater sensitivity to the larger cuts was observed and clear sensitivity to notches on the order of 0.63" was demonstrated. Note, a significant increase in the damage metric was observed after a two week delay between the end of the 0.25" notch cut and the start of the 0.31" notch cut. Relaxation of the boundary conditions over time was thought to be the source of the change. For validation, controlled time delays should be included into such studies to isolate and address long-time effects.

From the factor studies, the validation study design consisted of growing flaws by artificially cutting the structure at two fastener site locations. A series of environmental and boundary conditions were studied after each flaw growth scenario: temperature variation (+/- 40°F), temperature gradients, loading and unloading of 10 lb. mass, a simulated maintenance action at a set of fasteners including the case of minor loosening, and reinstallation and replacement of accelerometers. Much more detail on these factor studies and validation study design can be found in [15-16].

Model-based POD Analysis Approach

Conventional probability of detection (POD) evaluation for many quantitative NDE applications first uses empirical data to evaluate statistical relationships between the measurement response, $\hat{a}$, and the primary flaw size variable, *a*. Through application of a detection criterion as part of the NDE procedure, this statistical '$\hat{a}$ versus *a*' model can be used for evaluating the POD curve and probability of false call (POFC) rate, which together are usually referred to as "a POD model". The detection system can also be abstractly represented by a set of random variables a_i that act as inputs to a measurement model. Input variables can be categorized as being controlled (e.g. flaw size and material properties) or uncontrolled (e.g. liftoff, flaw morphology, and measurement noise). Detection consists of the measurement model output $\hat{a}$ being classified (or "called") according to pre-specified rules (e.g. a threshold).

The model-assisted POD (MAPOD) approach proposes to replace a conventional statistical fit in the measurement model with a complete physics-based model, *f*, calibrated for a given set of experimental conditions [27]. This relationship is given by:

$$\hat{a} = \beta_0 + \beta_1 f(a_i) + \varepsilon, \tag{2}$$

where β_0 and β_1 represent the model calibration parameters, and ε represents the residual error between the model and the experimental data. Estimating the statistics of β_0, β_1, and ε necessitates specific experimental sampling requirements. Variations due to flaw size and environmental (noise) conditions, for example, are represented in the model as probability distributions of the input variables. Hybrid models incorporating both empirical and physics-based components can be implemented to address all key factors including those that cannot be adequately simulated.

For this study, due to a lack of a validated physics-based model, a surrogate model fit using empirical data was developed for the evaluation. The primary variable associated with the critical flaw size is crack (notch) length, a_1. Controlled secondary variables in the study include flaw location (a_2), mean temperature (a_3), temperature gradients (a_4), ambient noise level (a_5). A response surface methodology was applied here to estimate the effect of each factor on the damage metric response and construct a model, $f(a_i)$ including uncertainty. Random events such as sensor failure/disbond (b_1), sensor bond degradation (b_2), sensor replacement (b_3), and local maintenance actions (b_4) were considered in the POD evaluation study. Assumptions concerning their frequency can be made and empirical models representing their effect can be evaluated and applied in conjunction with the scope of the SHM application.

To complete the POD evaluation, an assessment of the detection model under varying input conditions including uncertainty propagation is necessary. A second-order probabilistic approach has been developed to propagate both

aleatory uncertainty, due to inherent randomness in system behavior, and epistemic uncertainty, due to a lack of knowledge about values expected to be fixed [15-16]. Using this approach, epistemic variables are specified as intervals on values of parameters such as the means and standard deviations of random variables. For this study, the distributions for the input variables, mean temperature and temperature gradients, are presented in Fig. 4. Monte Carlo analysis is then applied here using outer and inner loops. The outer loop varies the values of distribution parameters of selected epistemic variables while the inner loop samples from the distributions.

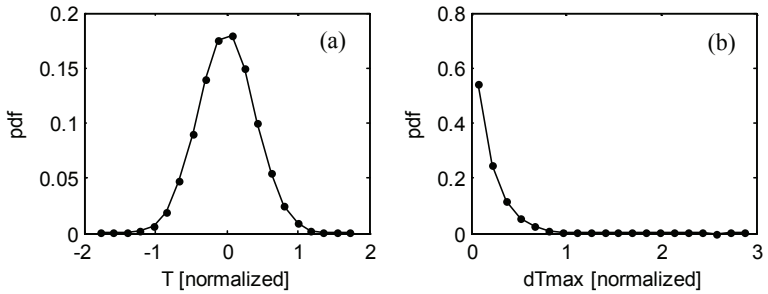

FIGURE 4. Parameter distributions: (a) mean normalized temperature (*T*), (b) maximum temperature gradient (*dTmax* / 10°F).

Sensitivity of POD to Flaw Location

Following acquisition of the experimental data, a regression model fit was performed using the R software environment. Three different flaw models were considered in the evaluation: (a) a flaw 2 and 3 combined evaluation (including flaw location factor, a_2), (b) a flaw 2 evaluation only, and (c) a flaw 3 evaluation only. One reason for performing and studying separate model fits for the different flaw growth sites was due to early observations that the SHM system was more sensitive to changes in flaw 3 with respect to changes in flaw 2. POD analysis results for the vibration-based SHM study are presented in Fig. 5 with respect to flaw size (in inches) for the case of a damage call threshold of 0.05. For each POD evaluation, both input parameter variation and model uncertainty are addressed through a two-level Monte Carlo simulation. From these results, there is clear need to separately evaluate the POD models for flaw 2 and flaw 3 locations. A single POD curve does not properly address the poor detection capability at the flaw 2 location as a function of performance at large flaw size. Using the 'flaw 2 and 3 combined' results will give one a false sense of security in terms of detection capability. Note, for any future SHM validation study, care must be taken to ensure the 'overall' POD capability evaluations do not mask 'isolated' flaw locations that have poor detection capability. Likewise, the low false call rate in the flaw 3 model is likely due to the flaw 3 model only including a portion of the simulation study variation.

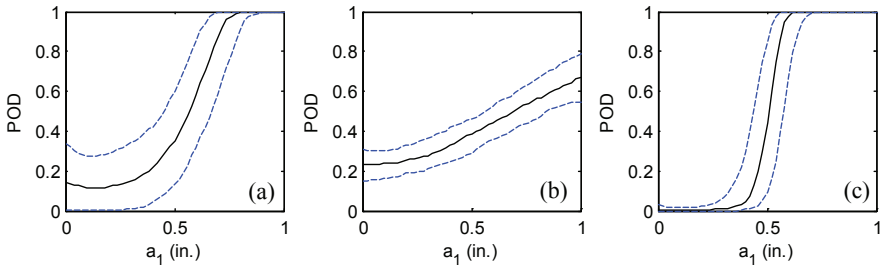

FIGURE 5. POD results with respect to flaw size including uncertainty bounds for (a) combined flaw 2 and 3, (b) flaw 2 only, and (c) flaw 3 only (*dm_threshold* = 0.05, use median sensor response).

Impact of Sensor Degradation over Time

This analysis approach also enables the evaluation of the impact of sensor durability on POD performance. Evidence from strain gauge sensor data on C-17 aircraft demonstrates the need for assessing the impact of degradation, where 22% of the sensors were infant failures and about 40% of the total failed within the first ten years of the aircraft

life [28]. Given that only eight sensors are present in the subject SHM system, the scenario of 25% failures, two accelerometers, was considered during the first 6 year period of operation. Fig. 6(a) presents two probability density functions for the time to failure for the first and second sensors. Data tables were constructed evaluating POD models for all of the 'single sensor' and 'two sensor' failure scenarios. A Monte Carlo simulation was then performed using 10000 samples from the time to failure distributions for the first and second sensors. Results are presented in Fig. 6(b) for the mean value from the composite Monte Carlo simulation POD results at a flaw size of 1.0 inch as a function of time. This analysis is useful because it highlights the sensitivity of certain flaw locations to degradation in the SHM system. In particular, the detection of flaw 2 suffers from weak crack sensitivity with respect to significant noise sensitivity due to varying temperature conditions.

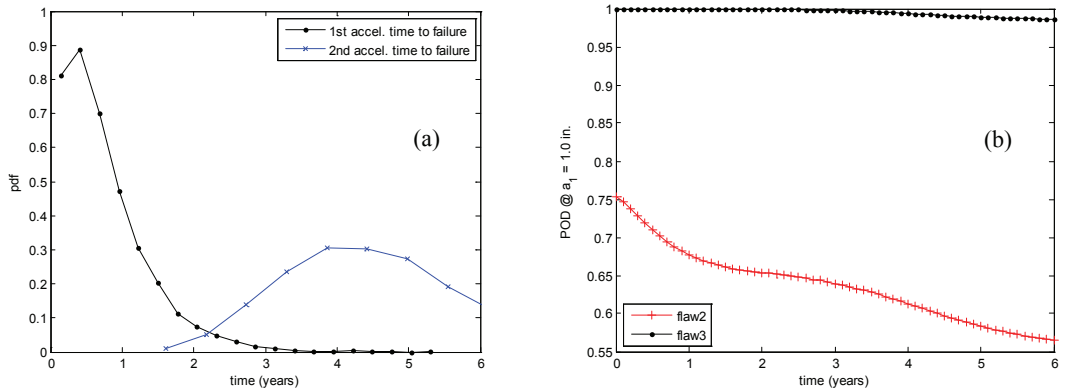

FIGURE 6. (a) Case study probability density functions for the time to failure for the first sensor and second sensor, (b) mean expected probability of detection (POD) at a flaw size of 1.0 in. with respect to time for all SHM systems found in the field.

CONCLUSIONS AND FUTURE WORK

To address validation of NDE characterization capability, a comprehensive approach to NDE characterization error evaluation was presented with a case study that follows the framework of the 'ahat-versus-a' model evaluation process for POD assessment. As well, work was presented on a protocol and demonstration for evaluating structural health monitoring system reliability. The design and results of the full validation study highlight the general protocol feasibility, emphasize the importance of evaluating the key application characteristics prior to the POD study, and demonstrate an approach to quantify varying sensor durability on the POD performance. However, challenges remain, in particular on how to properly address long time-scale effects with accelerated testing and how to address large testing requirements given the independence of each flaw location in the evaluation. More rigorous demonstrations are clearly needed for the most-promising SHM technologies. Model-assisted evaluation using validated NDE and SHM models should also help address these issues and mitigate the cost of performing these studies.

ACKNOWLEDGMENTS

This paper presents an overview of recent work on NDE and SHM characterization performance evaluation. This work has been supported through several contractual efforts with the U.S. Air Force Research Laboratory (AFRL): (1) through a SBIR Phase II contract awarded to Victor Technologies, LLC, AF112-130, FA8650-11-M-5180, (2) through a contract FA8650-09-C-5204 with Radiance Technologies, Inc., and (3) through the Research Initiatives for Materials State Sensing (RIMSS) contract FA8650-10-D-5210, to Computational Tools through Universal Technology Corporation (UTC). The authors would like to thank David Forsyth of TRI/Austin and Enrique Medina of UTC for discussions on NDE and SHM characterization metrics.

REFERENCES

1. U.S. Department of Defense, Standard Practice: Aircraft Structural Integrity Program (ASIP), MIL-STD-1530C (1 Nov 2005).
2. Lindgren, E. A., Knopp, J. S., Aldrin, J. C., Steffes, G. J., Buynak, C. F. "Aging Aircraft NDE: Capabilities, Challenges, And Opportunities", *Review of Progress in QNDE*, Vol. 26, AIP, pp. 1731-1738, (2007).

3. Medina, E. A. and Aldrin, J. C., "Value Assessment Approaches for Structural Life Management through SHM", *Encyclopedia of Structural Health Monitoring*, John Wiley & Sons, Ltd. (2009).
4. U.S. Department of Defense, Handbook, Nondestructive Evaluation System Reliability Assessment, MIL-HDBK-1823A, (7 April 2009).
5. Aldrin, J. C., Annis, C., Sabbagh, H. A., Knopp, J., and Lindgren, E. A., "Assessing the Reliability of Nondestructive Evaluation Methods for Damage Characterization," *Review of Progress in QNDE*, Vol. 33, AIP, pp. 2071-2080, (2014).
6. Nordtest Report, *Guidelines for NDE Reliability and Descriptions*, NT TECHN REPORT 394, Approved 1998-04.
7. Aldrin, J. C., Sabbagh, H. A., Annis, C., Shell, E. B., Knopp, J., and Lindgren, E. A., "Assessing Inversion Performance and Uncertainty in Eddy Current Crack Characterization Applications", *Review of Progress in QNDE*, Vol. 34, AIP, (2015).
8. Annis, C., Aldrin, J. C., and Sabbagh, H. A., "What is Missing in Nondestructive Testing Capability Evaluation?" *Materials Evaluation*, Vol. 73, n 1, pp. 44-54, (2015).
9. Kessler, S.S., "Certifying a Structural Health Monitoring System: Characterizing Durability, Reliability and Longevity," *Proceedings of the 1st International Forum on Integrated Systems Health Engineering and Management in Aerospace*, (Napa, CA, 7-10 November 2005).
10. Federal Aviation Administration, Advisory Circular, AC29-2C, DOT FAA, *Certification of transport category rotorcraft - rotorcraft health and usage monitoring systems (HUMS),* (2011).
11. SAE International, ARP-6461, *Guidelines for Implementation of Structural Health Monitoring on Fixed Wing Aircraft*, (2013).
12. Aldrin, J. C., Medina, E. A., Lindgren, E. A., Buynak, C. F., Steffes, G., Derriso, M., "Model-assisted probabilistic reliability assessment for structural health monitoring systems," *Review of Progress in QNDE*, Vol. 29, AIP, pp. 1965-1972, (2010).
13. Lindgren E.A., Buynak C.F., "The need and requirements for validating damage detection capability," In Chang FK, editor. *Proceedings of the 8th International Workshop on SHM*, Stanford: DEStech Publications, Inc. p. 2444-2451, (2011).
14. Aldrin J.C., Medina E.A., Lindgren E.A., Buynak C.F., Knopp J.S., "Protocol for reliability assessment of structural health monitoring systems incorporating model-assisted probability of detection (MAPOD) approach," In Chang FK, editor. *Proceedings of the 8th International Workshop on SHM*, Stanford: DEStech Publications, Inc. p. 2452-2459, (2011).
15. Medina E.A., Aldrin J.C., Santiago J., Lindgren E.A., Buynak C.F., "Demonstration of model assisted reliability assessment protocol on a proposal low frequency vibration based damage sensing case," In Chang FK, editor. *Proceedings of the 8th International Workshop on SHM*, Stanford: DEStech Publications, Inc. p. 2460-2467, (2011).
16. Aldrin, J. C., Medina, E. A., Santiago, J., Lindgren, E. A., Buynak, C., Knopp, J., "Demonstration study for reliability assessment of SHM systems incorporating model-assisted probability of detection approach," *Review of Progress in QNDE*, Vol. 31, AIP Conf. Proc. 1430, pp.1543-1550, (2012).
17. Brausch, J, and Steffes, G. "Demonstration, qualification, and airworthiness certification of structural damage sensing (SDS) systems for Air Force application," AFRL-RX-WP-TM-2013-0062, AFRL Report, Wright-Patterson AFB OH, (2013).
18. Gagar D, Irving P, Jennions I, Foote P, Read I, McFeat J., "Development of probability of detection structural health monitoring data for structural health monitoring damage detection techniques based on acoustic emission," In Chang F.K., editor. *Proceedings of the 8th International Workshop on SHM*, Stanford: DEStech Publications, Inc. p. 1391-1398, (2011).
19. Ihn, J. B., Pado, L., Leonard, M. S., Olson, S. E., & DeSimio, M. P., "Development and performance quantification of an ultrasonic structural health monitoring system for monitoring fatigue cracks on a complex aircraft structure," In Chang F.K., editor, *Proceedings of the 8th International Workshop on SHM*, Stanford: DesTech Publications, Inc., (2011).
20. Pado, L. E., J. B. Ihn, and J. P. Dunne, "Understanding probability of detection (POD) in structural health monitoring systems," In Chang F.K., editor, *Proceedings of the 9th International Workshop on SHM*, Stanford: DesTech Publications, Inc., (2013).
21. Mueller I, Janapati V, Banerjee S, Lonkar K, Roy S, Chang F.K.., "On the performance quantification of active sensing SHM systems using model-assisted POD methods," In Chang F.K., editor. *Proceedings of the 8th International Workshop on SHM*, Stanford: DesTech Publications, Inc. p. 2417-2428., (2011).
22. Schubert-Kabban A. C., Derriso M. M., "Certification in structural health monitoring systems," In Chang FK, editor. *Proceedings of the 8th International Workshop on SHM*, Stanford: DesTech Publications, Inc. p. 2429-2436, (2011).
23. Schubert-Kabban C., Greenwell B., DeSimio M., Derriso M., "The probability of detection for structural health monitoring systems: repeated measures data," *Structural Health Monitoring*, p. 1-13, (2015).
24. Kessler, S. S., Flynn, E. B., Dunn, C. T., and Todd, M. D., "A structural health monitoring software tool for optimization, diagnostics and prognostics," (Univ California San Diego, La Jolla, 2011).
25. Roach, D. "Validation and verification processes to certify SHM solutions for commercial aircraft applications," In Chang F.K., editor. *Proceedings of the 9th International Workshop on SHM*, Stanford: DesTech Publications, Inc., (2013).
26. 1st International Workshop on SHM Reliability, Chairs: Swindle, P. and Kessler, S., G-11 SHM (AISC-SHM) Technical Committee, (Cambridge, MA, April 13-14, 2015).
27. Thompson, R. B., "A unified approach to the model-assisted determination of probability of detection", *Materials Evaluation*, Vol. 66, pp. 667-673, (2008).
28. Aldrin, J. C., Knopp, J., Sabbagh, H. A., "Bayesian methods in probability of detection estimation and model-assisted probability of detection (MAPOD) evaluation," *Review of Progress in QNDE*, Vol. 32, AIP, pp.1733-1740, (2013).
29. Ware, R., Reams, R., Woods, A., Selder, R., "Sensor Reliability in Fielded C-17 Aircraft Strain Gauges," *Proceedings of the 4th International Workshop on Structural Health Monitoring*, Stanford, CA, pp. 478-486, (2005).